

ANAND KAMBLE

San Francisco, CA | +1 (448) 200-9180 | anandmk837@gmail.com | [LinkedIn](#)

Work Experience

Liebre.ai, *Founding Engineer*

May 2025 – Present

- Core engineer on Luca, an AI accounting agent automating double-entry bookkeeping — built its graph-based workflow engine (~40 workflows, ~50 skills) on FastAPI, PostgreSQL, and Redis.
- Designed a transaction-clustering service (text embeddings + UMAP + categorical flags) with ~95% cluster purity.
- Built LLM abstraction layers with versioned prompt management and OpenTelemetry tracing; deployed LLM/ML microservices improving Time To First Token.
- Built NightShift from scratch — an agent harness running fleets of autonomous Claude Code agents on ClickUp tickets through worker, QA, and PR-review stages 24/7: custom system prompts, lifecycle hooks, and hook-driven session-state detection over PTY-backed terminals (TypeScript, React, xterm.js).
- Built real-time agent streaming: SSE/WebSocket events, a Redis-backed event bus, and JSON Patch state sync.

Florida State University, *Graduate Research Assistant*

Jan 2024 – Oct 2024

- Designed and implemented retrieval-augmented generation pipelines, improving accuracy of applications.
- Built secure REST APIs to expose LLM predictions for downstream applications.
- Integrated multiple LLM APIs and self-hosted models for flexible inference.
- Evaluated model performance on synthetic datasets and iterated on prompt engineering.
- Orchestrated multi-agent workflows to automate complex task sequences.
- Designed and deployed model-parallel inference infrastructure on the university's SLURM-managed HPC cluster, achieving 8x speedup.

IncerroAI, *Frontend Engineer*

Apr 2022 – Nov 2022

- Developed high-performance web applications featuring real-time data binding, and responsive layouts.
- Optimized React rendering performance, reducing re-renders by 40% through memoization and efficient state management.

Education

Florida State University

Master of Science, Computational Science

Aug 2023 – Dec 2024

- Coursework: Machine Learning, AI Methods, Data Mining, High-Performance Computing, Monte Carlo Methods

Skills

AI Agents & Harness Engineering: Agent architecture & orchestration, Claude Code / pi harnesses (hooks, subagents, custom system prompts), Model Context Protocol (MCP), autonomous multi-agent pipelines

Programming Languages: Python, TypeScript, Java, C/C++, Bash, SQL

Frameworks & Libraries: LangGraph, LangChain, FastAPI, LlamaIndex, llama.cpp, LiteLLM, PyTorch, scikit-learn, React, Express

Testing: PyTest, Vitest, Playwright, Testcontainers

Cloud & MLOps: GCP (Cloud Run, Cloud SQL), AWS, Docker, Kubernetes, GitHub Actions, Redis, PostgreSQL, SLURM

Tools: Git, Linux/Unix, OpenTelemetry, JIRA

Publications & Preprints

[Google Scholar](#) · 26 citations, h-index 4

- [NAVER LABS System Re-implementation for the IWSLT 2026 Instruction-Following Task](#). arXiv:2607.05623, 2026.
- [Transcription and translation of videos using fine-tuned XLSR Wav2Vec2 on custom dataset and mBART](#). arXiv:2403.00212, 2024.
- [End to end Hindi to English speech conversion using Bark, mBART and a finetuned XLSR Wav2Vec2](#). arXiv:2401.06183, 2024.
- [Custom Data Augmentation for low resource ASR using Bark and Retrieval-Based Voice Conversion](#). arXiv:2311.14836, 2023.
- [Object following robot based on AI/ML](#). *Materials Today: Proceedings* (Elsevier), 2022.

Projects

Open Source – MCP Instagram ([github](#))

- MCP server enabling AI assistants to interact with Instagram via the Model Context Protocol.

Automatic Grading System using LLM

- Developed an AI-powered grading platform leveraging fine-tuned models and RAG to assess assignments.
- Improved grading accuracy through advanced prompt engineering and reduced hallucinations.

Aircraft Fault ATA Prediction

- Fine-tuned Phi-3-mini model on aircraft maintenance logs to predict ATA codes, leveraging LoRA adapters, 8-bit mixed-precision.
- Built a configurable pipeline that can process new logs and assign correct ATA codes.